Predictive models avoid excessive reductionism in cognitive neuroimaging

Gaël Varoquaux[a,b,*], Russell A. Poldrack[c]

[a] *Parietal project-team, Inria Saclay-île de France, France*
[b] *CEA/Neurospin bât 145, 91191 Gif-Sur-Yvette, France*
[c] *Department of Psychology, Stanford University, Stanford, California, USA*

Understanding the organization of complex behavior as it relates to the brain requires modeling the behavior, the relevant mental processes, and the corresponding neural activity. Experiments in cognitive neuroscience typically study a psychological process via controlled manipulations, reducing behavior to one of its component. Such reductionism can easily lead to paradigm-bound theories. Predictive models can generalize brain-mind associations to arbitrary new tasks and stimuli. We argue that they are needed to broaden theories beyond specific paradigms. Predicting behavior from neural activity can support robust reverse inference, isolating brain structures that support particular mental processes. The converse prediction enables modeling brain responses as a function of a complete description of the task, rather than building on oppositions.

Keywords: machine learning, cognitive models, psychology, brain imaging

Perception and action build upon an array of mental processes, which have been characterized in detail by the psychological sciences. However, unifying these psychological processes to model the relation between brain and mind in any given situation remains a great challenge (Newell, 1973•). The principal challenge lies in finding the appropriate representation of the components of cognition and behavior. Cognitive neuroscience builds upon controlled experiments to isolate these components and link them to brain activity. But quantifying the effects of any manipulation requires that the psychological or behavioral components of interest be clearly specified (Poldrack and Yarkoni, 2016; Krakauer et al., 2017•). Isolating components of mental processing leads to studying them only via oppositions, and this reductionism prevents the building of broad theories of the mind.

We believe that predictive modeling provides new tools to tackle this formidable task. The accumulation of a broad range of shared data in cognitive neuroimaging provides a widely varied set of observations of brain and behavior. Using these data, models can be built that accurately describe multiple experiments, going beyond the surface description of tasks to identify associations between brain systems and underlying mental processes that span across tasks. Whereas cognitive neuroscience has typically focused on particular theoretical oppositions, this approach instead builds models that generalize beyond specific tasks, based on the methodology of machine learning with a focus on out-of-sample prediction rather than the detection of specific experimental effects (Yarkoni and Westfall, 2017).

Data-driven approaches are often distinguished from hypothesis-driven research, with the implication that data-driven work is necessarily theory-free. However, we argue that data-intensive methods can actually provide the basis for building broader theories, which abstract away from the specifics of any particular experimental approach and thus have the potential to generalize to a much larger range of phenomena.

Prediction allows generalization to arbitrary new tasks and situations

Predictive models give a specific prediction, forecasting a target quantity numerical or categorical based on new data. In the framework of statistical machine learning, they contain tunable parameters that are adjusted to learn the association between data and target, usually with some penalty for model complexity (James et al., 2013). Performance can then be tested on unseen data. Predictive models provide powerful tools to learn brain-mind associations from recordings of neural activity (Varoquaux and Thirion, 2014; Pereira et al., 2009; Lebedev and Nicolelis, 2006). Compared to traditional computational neuroscience models, they allow generalization to new tasks and situations. Hence, they can provide conclusions that are not paradigm-bound, as cognitive neuroscience largely has been.

Broad generalizations can bridge very different situations. For instance Knops et al. (2009) showed that learning brain-activity patterns that discriminate right from left saccades would classify mental additions as rightward saccades and subtractions as leftward. Whether or not brain-mind associations generalize can be tested across paradigms, and such tests highlight universal components of psychological processes. Studying memory encoding,

*Corresponding author

Predictive models versus classic neuroimaging analyses
By predictive models we refer to mathematical models of the data that explicitly minimize prediction error, typically relying on methods from machine learning. They differ from standard neuroimaging statistical methods in several way. First, their validity is established by successful predictions from new data, and not by isolating significant differences across observations. This test is sound without the need for modeling assumption, such as Gaussian noise. As a consequence, predictive models often are more complex, departing from maximum-likelihood estimates and fitting more unknowns variables than available observations (Wu et al., 2006; James et al., 2013).
For neuroimaging, this additional complexity opens the door to modeling more effects jointly, and therefore building models that describe more than an isolated dimension of cognition. Also, given that a model is validated by its predictions, rather than by its goodness of fit on the data used to fit the model, validation can be done using different constructs than those used to build the model, for instance using high-level descriptions of cognition instead of features of the experiments or brain signals. Testing for out-of-sample properties, as when measuring prediction, helps avoiding some pitfalls of in-sample model testing using p-values, such as circularity or trivial effects (Wasserstein et al., 2016).

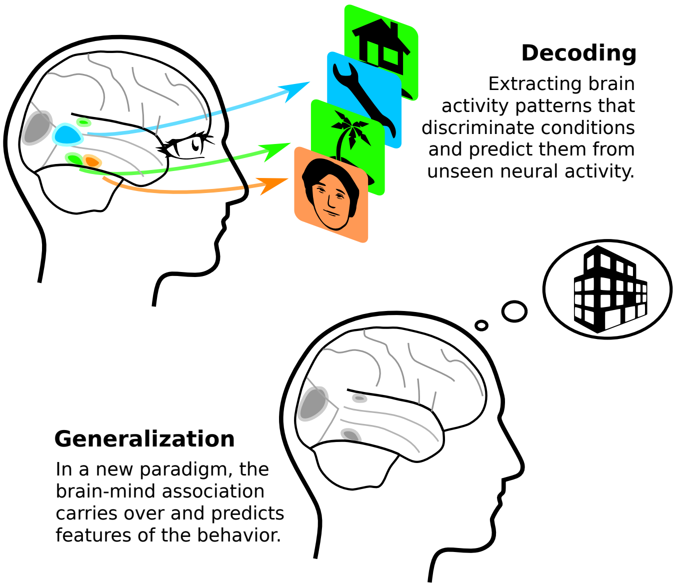

Figure 1: **An example of brain decoding and generalization**: Brain decoding finds patterns in brain-activity data that discriminate given conditions on unseen data, here identifying perceptual stimuli. The corresponding brain-mind association can then be generalized to a new paradigm, or new stimuli, *eg* by applying the decoder in a visual memory task.

Polyn et al. (2005) used a classifier to generalize from perception to memory retrieval to show that both processes share a common neural substrate. In research on pain, neural evidence supports separating physical and emotional pain, as brain activity can robustly discriminate them across studies (Wager et al., 2013••).

Generalization to arbitrary tasks supports reverse inference

Associating neural activity in a brain structure or network to a predicted behavior that generalizes to arbitrary tasks provides a deeper understanding of the computations supported by the structure. Bound to a given paradigm, such characterization would be an invalid reverse inference, as in an experiment only neural activity is observed as a consequence of a psychological manipulation, rather than causing a mental state (Poldrack, 2006). Decoding studies, which predict behavior from observed activity, provide evidence for reverse inference. However, to characterize function beyond simple oppositions, decoding must be applied across a very broad sampling of behavior and cognition. Analyzing many different studies jointly provides a practical solution to sample such a variety of mental states (Poldrack, 2011•). Generalization across experimental paradigms shows that identified brain structures are not a mere consequence of experimental details of the task (Varoquaux et al., 2018).

Causal questions are central to interpreting results of a neuroimaging experiment: what in the observed brain activity is a consequence of the task, versus a cause of the behavior or the mental state? Comparing which brain structures are activated in a standard analysis to which ones support decoding can rule out structures that mediate a function but do not cause it. In stimuli-driven experiments, detecting a structure in both models suggests that it is a direct consequence of the stimulus, for instance activity in the fusiform face area for face recognition, rather than a non-specific side effect, such as activity in primary visual areas. Such side effect cannot support decoding (Weichwald et al., 2015).

Formal representations of tasks and behavior are important

Modeling brain activity beyond a specific paradigm needs robust and general descriptions of stimuli, tasks, and behavior (Turner and Laird, 2012; Poldrack et al., 2011b). Building such descriptions faces two challenges: capturing all the relevant mental processes in a task, and formalizing their relationships. The psychological manipulations of a task provide information regarding the main mental processes it recruits. Yet, it is necessary to go beyond the primary effect of interest: a visual n-back experiment is not only a memory experiment, but also involves visual processing ,decision making, and many other functions (Varoquaux et al., 2018). Relating mental processes across studies raises many subtle questions, for instance whether to distinguish between autobiographical and episodic memory. Taxonomies or ontologies give a formal framework to capture this knowledge (Poldrack and Yarkoni, 2016).

However, using cognitive ontologies in brain mapping is a chicken-and-egg problem, given that neuroimaging should also inform these ontologies, via the links between mental processes that it reveals. Data-driven semantic techniques hold promise to build representations of cognitive neuroscience concepts informed by brain data (Poldrack et al., 2012•;Yeo et al., 2015; Bolt et al., 2017). Used across the literature, they can overcome variations in terminology and link mental processes that elicit similar activity or build more open-ended encoding or decoding models (Yarkoni et al., 2011; Rubin et al., 2017; Dockès et al., 2018).

Encoding models extract better representations of tasks

Encoding models, which predict brain data from the task (Naselaris et al., 2011•), can go beyond describing a task with a small set of labels and capture details and interplay in its components. Because they allow for the use of much richer and less constrained descriptions of tasks (Wu et al., 2006), they can ground a model of brain function beyond specific experimental paradigms. To model psychological manipulations, encoding studies usually rely on machine-learning techniques that are well suited to complex and high-dimensional representations. Their testing procedure measures how well a representation of the stimuli or task can predict brain activity on unseen data, unlike standard analyses in brain imaging which use general linear models to detect significant effects of oppositions in brain activity (Poldrack et al., 2011a).

Encoding models replace crafting stimuli

Progress in visual neuroscience provides a striking example of the success of encoding models. Historically, our understanding of the human visual system has been driven by experiments crafting stimuli to reveal the selectivity of a hierarchy of brain modules (Grill-Spector and Malach, 2004), from spots and slab that revealed edge detectors in the visual area V1 (Hubel and Wiesel, 1959) to more complex shapes mapping mid-level regions (Logothetis et al., 1995), to semantic regions, studied via stimuli such as faces or scenes (Haxby et al., 2001; Kanwisher et al., 1997). These experiments have been very successful in providing a conceptual model of the human visual system, but no one single experiment could ground this model of visual processing. Rather, it relied on combining interpretations across studies of disparate and non-ecological stimuli.

On more ecological stimuli, rich encoding models can reveal the properties of the primary visual cortex (Kay et al., 2008; Miyawaki et al., 2008), transforming them into a representation that provides the ability to accurately predict brain responses to complex stimuli (Naselaris et al., 2011•). To go beyond primary areas, rich statistical models of natural images are needed. Deep artificial neural networks developed for computer vision build representations of the images used as stimuli that accurately predict brain responses, outperforming computational-neuroscience models of vision to explain the workings of mid-level areas (Khaligh-Razavi and Kriegeskorte, 2014; Yamins et al., 2014••). They build a hierarchy of intermediate representations, from low-level edge detectors to more semantic information, that maps well to the full hierarchy of human visual processing (Güçlü and van Gerven, 2015; Eickenberg et al., 2017••). These results extend to high-level areas the hypothesis that V1 is tuned to Gabor filters because these form good statistical representations of natural images (Olshausen et al., 1996).

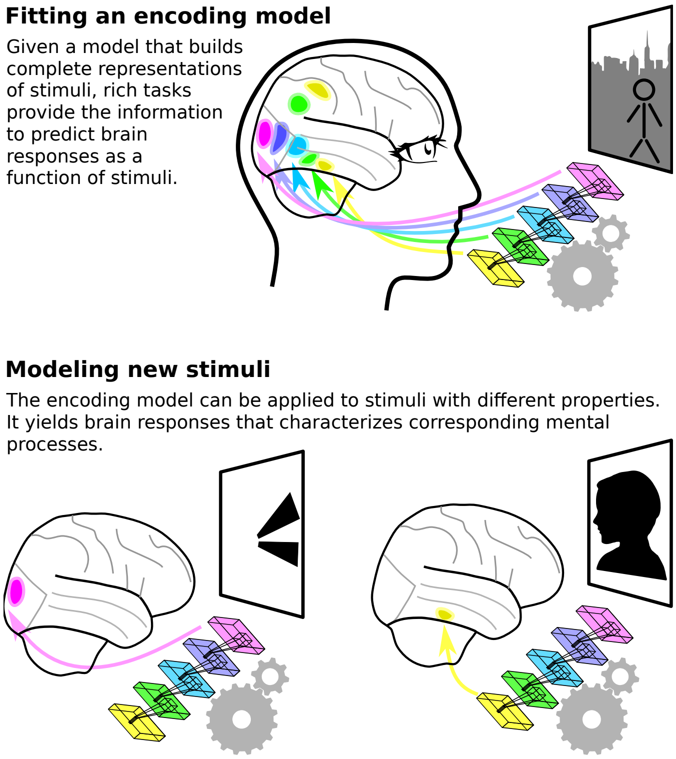

Figure 2: **Encoding methodology and generalization**: Encoding models capture the full details of brain responses given a rich description of the stimuli or tasks. This behavior-to-brain association can then be generalized to new stimuli or tasks, to characterize mental processes.

The latest studies do not require hand crafting stimuli to elicit controlled responses in a given region. They provide a full mapping of the computational steps of human vision from natural stimuli, and they jointly model all stimulus features important to brain responses, such that their model of brain response can generalize beyond the paradigm used in any given experiment. For example, Eickenberg et al. (2017)•• showed that a model learned from subjects watching natural images could generate both retinotopic maps and face versus scene contrasts.

The study of other sensory systems has demonstrated similar success. Encoding models have mapped the functional modules of the auditory cortex from spectrograms (Santoro et al., 2017), or using a hierarchy of representations isolated by deep neural networks used to process sounds on computers (Kell et al., 2018). Perceptual sciences lend themselves well to rich encoding models: as neuroimaging experiments can use rapid successions of trials with stimuli that are easily characterized across a broad set of features, providing large data accumulation.

Precise models of responses to stimuli capture high-level processes

Data-intensive encoding models can also map more complex psychological functions. In particular, encoding models can accurately capture semantic representations of language in the brain at the level of words (Mitchell et al., 2008), short texts (Wehbe et al., 2014), and stories (Huth et al., 2016). Finely-tuned models of stimuli responses provide excellent windows to attentional (Çukur et al., 2013; Hausfeld et al., 2018) or perceptual decision (Gwilliams and King, 2017) mechanisms. They relate better to experimental data than more conceptual models such as drift diffusion models (Gwilliams and King, 2017). Finally, using complex stimuli, such as a full movie, enables an ecological study of processes such as episodic memory (Baldassano et al., 2017•).

In all these settings, predictive modeling using encoding models enables the study of sensory and cognitive processes without reducing the experiment to one simple opposition or variation. As a result, predictions hold across paradigms and generalize to novel paradigms (Eickenberg et al., 2017••).

Prediction is a guiding principle to modeling stimuli

A key challenge for modeling in cognitive neuroscience is to generate rich representations of stimuli. Given data and a task to perform on it, artificial intelligence models extract representations that are optimal for information processing. These representations provide a good basis to study brain responses because they give a complete view of the task, but also because both artificial intelligence and human cognition capture the statistical regularities of our world. To model stimuli, minimizing prediction error is a guiding principle to extract relevant representations. For example, prediction of neighboring words is central to both cognitive and computational linguistic (Willems et al., 2016). Artificial neural networks optimized for object recognition form good representations to study object recognition in the ventral stream (Yamins and DiCarlo, 2016••). Extending beyond prediction or recognition, recent progress in machine learning synthesizes rich data such as images guided by models that discriminate real from fake images (Goodfellow et al., 2014).

Generalization will ground broader theories

Scientific endeavors strive for conclusions that generalize, predicting features of new situations. While cognitive neuroscience has often focused on informal generalizations, machine-learning techniques will bring more precise predictions and more general models. Generalization across paradigms is key to achieving broader theories, in decoding to isolate the neural supports of mental states or in encoding to build complete descriptions of behavior.

References

Baldassano, C., Chen, J., Zadbood, A., Pillow, J.W., Hasson, U., Norman, K.A., 2017. Discovering event structure in continuous narrative perception and memory. Neuron 95, 709–721.e5.
• This work studies memory creation and consolidation via a data-driven segmentation of neural responses to movie watching into nested events.

Bolt, T., Nomi, J.S., Yeo, B.T.T., Uddin, L.Q., 2017. Data-Driven extraction of a nested model of human brain function. J. Neurosci. 37, 7263–7277.

Çukur, T., Nishimoto, S., Huth, A.G., Gallant, J.L., 2013. Attention during natural vision warps semantic representation across the human brain. Nat. Neurosci. 16, 763–770.

Dockès, J., Wassermann, D., Poldrack, R., Suchanek, F., Thirion, B., Varoquaux, G., 2018. Text to brain: predicting the spatial distribution of neuroimaging observations from text reports. MICCAI .

Eickenberg, M., Gramfort, A., Varoquaux, G., Thirion, B., 2017. Seeing it all: Convolutional network layers map the function of the human visual system. NeuroImage 152, 184–194.
•• This study shows that artificial neural networks used for computer vision can capture the neural responses to visual stimuli with a association that can then be used to distinguish functional modules in the full visual cortex, or to predict the neural responses to classic paradigms, such as retinotopy or face-scene oppositions.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative adversarial nets, in: Neural information processing systems, p. 2672.

Grill-Spector, K., Malach, R., 2004. The human visual cortex. Annu. Rev. Neurosci. 27, 649.

Güçlü, U., van Gerven, M.A., 2015. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. Journal of Neuroscience 35, 10005–10014.

Gwilliams, L., King, J.R., 2017. Performance-optimized hierarchical models only partially predict neural responses during perceptual decision making. bioRxiv , 221630.

Hausfeld, L., Riecke, L., Formisano, E., 2018. Acoustic and higher-level representations of naturalistic auditory scenes in human auditory and frontal cortex. Neuroimage 173, 472–483.

Haxby, J.V., Gobbini, I.M., Furey, M.L., Ishai, A., Schouten, J.L., Pietrini, P., 2001. Distributed and overlapping representations of faces and objects in ventral temporal cortex. Science 293, 2425.

Hubel, D.H., Wiesel, T.N., 1959. Receptive fields of single neurones in the cat's striate cortex. J. Physiol. 148, 574–591.

Huth, A.G., de Heer, W.A., Griffiths, T.L., Theunissen, F.E., Gallant, J.L., 2016. Natural speech reveals the semantic maps that tile human cerebral cortex. Nature 532, 453–458.

James, G., Witten, D., Hastie, T., Tibshirani, R., 2013. An introduction to statistical learning. volume 112. Springer.

Kanwisher, N., McDermott, J., Chun, M.M., 1997. The fusiform face area: a module in human extrastriate cortex specialized for face perception. J. Neurosci. 17, 4302–4311.

Kay, K.N., Naselaris, T., Prenger, R.J., Gallant, J.L., 2008. Identifying natural images from human brain activity. Nature 452, 352.

Kell, A.J.E., Yamins, D.L.K., Shook, E.N., Norman-Haignere, S.V., McDermott, J.H., 2018. A Task-Optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. Neuron .

Khaligh-Razavi, S.M., Kriegeskorte, N., 2014. Deep supervised, but not unsupervised, models may explain IT cortical representation. PLoS Comput. Biol. 10, e1003915.

Knops, A., Thirion, B., Hubbard, E., Michel, V., Dehaene, S., 2009. Recruitment of an area involved in eye movements during mental arithmetic. Science 324, 1583.

Krakauer, J.W., Ghazanfar, A.A., Gomez-Marin, A., MacIver, M.A., Poeppel, D., 2017. Neuroscience needs behavior: correcting a reductionist bias. Neuron 93, 480–490.

• Makes the case that, in addition to detailed models of neural activity, decomposition of behavior is needed to understand brain function, hence careful behavioral work must complement neural manipulations.

Lebedev, M.A., Nicolelis, M.A., 2006. Brain–machine interfaces: past, present and future. TRENDS in Neurosciences 29, 536–546.

Logothetis, N.K., Pauls, J., Poggio, T., 1995. Shape representation in the inferior temporal cortex of monkeys. Current Biology 5, 552.

Mitchell, T.M., Shinkareva, S.V., Carlson, A., Chang, K.M., Malave, V.L., Mason, R.A., Just, M.A., 2008. Predicting human brain activity associated with the meanings of nouns. science 320, 1191.

Miyawaki, Y., Uchida, H., Yamashita, O., Sato, M.A., Morito, Y., Tanabe, H.C., Sadato, N., Kamitani, Y., 2008. Visual image reconstruction from human brain activity using a combination of multiscale local image decoders. Neuron 60, 915–929.

Naselaris, T., Kay, K.N., Nishimoto, S., Gallant, J.L., 2011. Encoding and decoding in fMRI. Neuroimage 56, 400.

• This paper lays out a methodology to identify brain-behavior associations combining encoding and decoding, where encoding is used with a representational model of behavior, and decoding shows clear evidence of the association.

Newell, A., 1973. You can't play 20 questions with nature and win: Projective comments on the papers of this symposium

• Highlights the difficulty of giving an integrated picture on results on different mental processes involved in vision.

Olshausen, B., et al., 1996. Emergence of simple-cell remainsceptive field properties by learning a sparse code for natural images. Nature 381, 607.

Pereira, F., Mitchell, T., Botvinick, M., 2009. Machine learning classifiers and fMRI: a tutorial overview. Neuroimage 45, S199.

Poldrack, R., 2006. Can cognitive processes be inferred from neuroimaging data? Trends in cognitive sciences 10, 59.

Poldrack, R., Mumford, J., Nichols, T., 2011a. Handbook of functional MRI data analysis. University Press, Cambridge.

Poldrack, R.A., 2011. Inferring mental states from neuroimaging data: from reverse inference to large-scale decoding. Neuron 72, 692.

• This review show how decoding can be used on large-scale databases to ground formal reverse inference, capturing how selectively a brain area is activated by a mental process.

Poldrack, R.A., Kittur, A., Kalar, D., Miller, E., Seppa, C., Gil, Y., Parker, D.S., Sabb, F.W., Bilder, R.M., 2011b. The cognitive atlas: toward a knowledge foundation for cognitive neuroscience. Frontiers in neuroinformatics 5, 17.

Poldrack, R.A., Mumford, J.A., Schonberg, T., Kalar, D., Barman, B., Yarkoni, T., 2012. Discovering relations between mind, brain, and mental disorders using topic mapping. PLoS computational biology 8, e1002707.

• This study uses topic-extraction techniques applied to a large database of neuroimaging results to identify the conceptual structure of mental functions and their mapping to brain systems and the relations between mental disorders.

Poldrack, R.A., Yarkoni, T., 2016. From brain maps to cognitive ontologies: Informatics and the search for mental structure. Annual Review of Psychology 67, 587–612.

Polyn, S.M., Natu, V.S., Cohen, J.D., Norman, K.A., 2005. Category-specific cortical activity precedes retrieval during memory search. Science 310, 1963–1966.

Rubin, T.N., Koyejo, O., Gorgolewski, K.J., Jones, M.N., Poldrack, R.A., Yarkoni, T., 2017. Decoding brain activity using a large-scale probabilistic functional-anatomical atlas of human cognition. PLoS Comput. Biol. 13, e1005649.

Santoro, R., Moerel, M., De Martino, F., Valente, G., Ugurbil, K., Yacoub, E., Formisano, E., 2017. Reconstructing the spectrotemporal modulations of real-life sounds from fMRI response patterns. Proc. Natl. Acad. Sci. U. S. A. 114, 4799–4804.

Turner, J.A., Laird, A.R., 2012. The cognitive paradigm ontology: design and application. Neuroinformatics 10, 57–66.

Varoquaux, G., Schwartz, Y., Poldrack, R.A., Gauthier, B., Bzdok, D., Poline, J.B., Thirion, B., 2018. Atlases of cognition with large-scale human brain mapping. PLOS Computational Biology 14, 1–18.

Varoquaux, G., Thirion, B., 2014. How machine learning is shaping cognitive neuroimaging. GigaScience 3, 28.

Wager, T.D., Atlas, L.Y., Lindquist, M.A., Roy, M., Woo, C.W., Kross, E., 2013. An fMRI-based neurologic signature of physical pain. New England Journal of Medicine 368, 1388.

•• This study shows that fMRI can predict the level of physical pain in a way that is specific to physical pain, as opposed to social pain. The study relies on predictive models used across several pain-related experiments.

Wasserstein, R.L., Lazar, N.A., et al., 2016. The asas statement on p-values: context, process, and purpose. The American Statistician 70, 129–133.

Wehbe, L., Murphy, B., Talukdar, P., Fyshe, A., Ramdas, A., Mitchell, T., 2014. Simultaneously uncovering the patterns of brain regions involved in different story reading subprocesses. PLoS One 9, e112575.

Weichwald, S., Meyer, T., Özdenizci, O., Schölkopf, B., Ball, T., Grosse-Wentrup, M., 2015. Causal interpretation rules for encoding and decoding models in neuroimaging. NeuroImage 110, 48–59.

Willems, R.M., Frank, S.L., Nijhof, A.D., Hagoort, P., van den Bosch, A., 2016. Prediction during natural language comprehension. Cereb. Cortex 26, 2506–2516.

Wu, M.C.K., David, S.V., Gallant, J.L., 2006. Complete functional characterization of sensory neurons by system identification. Annu. Rev. Neurosci. 29, 477–505.

Yamins, D.L., Hong, H., Cadieu, C.F., Solomon, E.A., Seibert, D., DiCarlo, J.J., 2014. Performance-optimized hierarchical models predict neural responses in higher visual cortex. Proc Natl Acad Sci , 201403112

•• This study shows that models of neural response based on computer-vision artificial networks explain brain activity better than classic theoretical-neuroscience models of vision.

Yamins, D.L.K., DiCarlo, J.J., 2016. Using goal-driven deep learning models to understand sensory cortex. Nat. Neurosci. 19, 356–365.

•• Reviews the uses of artificial neural networks, developed for computer vision and artificial intelligence communities, to model neural responses in higher visual cortical areas.

Yarkoni, T., Poldrack, R.A., Nichols, T.E., Van Essen, D.C., Wager, T.D., 2011. Large-scale automated synthesis of human functional neuroimaging data. Nature methods 8, 665.

Yarkoni, T., Westfall, J., 2017. Choosing prediction over explanation in psychology: Lessons from machine learning. Perspectives on Psychological Science 12, 1100–1122.

•• This paper argues that psychology's focus on explaining the causes of behavior provides theories with little ability to predict future behaviors with any appreciable accuracy. It proposes to use principles from machine learning for greater understanding of behavior.

Yeo, B.T.T., Krienen, F.M., Eickhoff, S.B., Yaakub, S.N., Fox, P.T., Buckner, R.L., Asplund, C.L., Chee, M.W.L., 2015. Functional specialization and flexibility in human association cortex. Cereb. Cortex 25, 3654–3672.